A DATA BLUEPRINT FOR ADJUDICATING AI DEBATE

Gefei Liu* **Sonya Rashkovan*** **Sophia Lloyd George** **Isaac Sheidlower** **Serena Booth**
Brown University, Providence, RI 02912, USA
*Equal contribution

1 INTRODUCTION

AI systems are increasingly trained and evaluated using human feedback. A dominant paradigm for scalable supervision is single-turn preference labeling and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). While effective in many settings, single-turn adjudication provides a narrow interaction structure that can introduce limitations such as emphasis on surface-level presentation and over-optimization toward proxy reward signals (Ziegler et al., 2020; Moskovitz et al., 2023). An alternative proposal is to train and evaluate systems through *debate*, in which agents present competing arguments across multiple turns until a judge has sufficient information to adjudicate in favor of truth (Irving et al., 2018). Debate has promise for improving AI's capacity for reasoning and adherence to factuality compared to single-model baselines (Du et al., 2023) and to help humans reach more accurate conclusions when guided by competing expert arguments (Khan et al., 2024). However, debate is only as reliable as the adjudication process that converts an interaction into a supervision signal.

The promise of AI debate relies on the assumption that adjudication approximates an ideal and reliable decision-maker, in which the judge adjudicates in favor of truthful arguments. In practice, human judges exhibit substantial variability and systematic biases. Human adjudication is shaped by cognitive heuristics and can exhibit high variance even in high-stakes domains with extensive requisite training such as legal decision-making (Guthrie et al., 2007; 2002). These findings motivate the need to explicitly model the adjudication process to account for these biases rather than treating this supervision signal as derived from a perfect oracle.

Progress toward realistic adjudication models is limited by data. Existing debate datasets capture valuable subsets of dialogue (Zhang et al., 2016; Tiwari et al., 2025; Ruiz-Dolz & Iranzo-Sánchez, 2024; Abbott et al., 2016), evidence (Roush et al., 2024; Abbott et al., 2016; Greschner et al., 2025), or outcomes (Zhang et al., 2016; Greschner et al., 2025; Tiwari et al., 2025), but rarely record these components jointly, making it difficult to model how variations in debate components influence adjudications. This paper presents a blueprint for collecting structured datasets of debate rounds that couple turn-structured dialogue, evidence grounding, and adjudication traces (Sec. 2), along with collection protocols for obtaining such data (Sec. 3). Looking forward, such carefully curated datasets support the development of robust training and evaluation protocols for debate adjudication grounded in realistic human judgment processes (Sec. 4).

2 DATA SCHEMA

Modeling human adjudication requires datasets that holistically capture the full dynamics of a debate round: what arguments are exchanged, how evidence is weighed, and how judges ultimately arrive at decisions. Adjudication is a complex cognitive and social process shaped by the structure of interaction itself. Judges update their beliefs across turns, rely on heuristics when evaluating competing claims, and respond to persuasion strategies as well as substance. Outcomes can depend on features such as verbosity, confidence, or emotional framing. Our proposed approach is therefore divided into the following threefold structure.

Debate Dialogues Debate dialogue data should capture the full interaction structure of a debate round because adjudication unfolds over the course of the round (Zhang et al., 2016; Tiwari et al., 2025; Ruiz-Dolz & Iranzo-Sánchez, 2024; Abbott et al., 2016). Recording the debate *motion* and *summary* situates arguments within a debate context, while identifying *participants* and *roles* allows analysis of how credibility, expertise, and speaking position influence judgment. A complete, ordered transcript of *paragraphs* linked to speakers, roles, and debate stages enables modeling of belief updates across *turns* and the persuasive dynamics of interaction. Additional *annotations* such as interruptions or laughter help capture conversational cues that shape perceived confidence and persuasiveness. Together, these variables make it possible to study how interaction structure affects adjudication outcomes.

Evidence Grounding Evidence grounding must be recorded because judges also evaluate how arguments are supported in a debate (Roush et al., 2024; Abbott et al., 2016; Greschner et al., 2025).

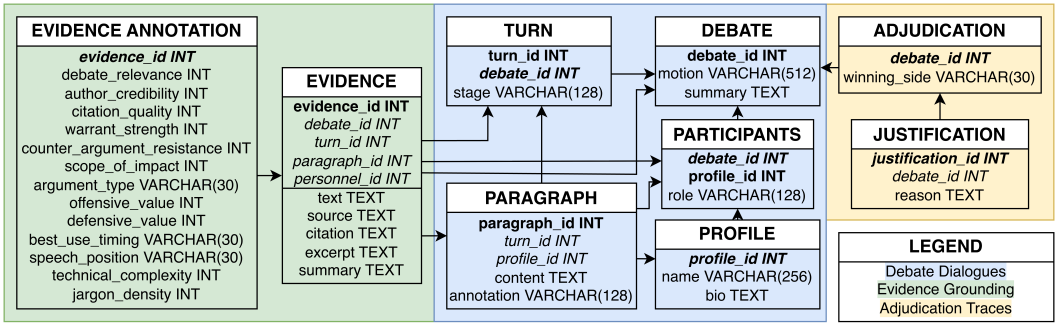

Figure 1: Relational schema for a debate adjudication dataset, capturing debate dialogue, evidence grounding, and adjudication traces. Listed variables are illustrative examples that may be collected. Arrows indicate contribution relationships; **bold** denotes primary keys and *italics* denotes foreign keys.

Linking each piece of evidence to its debate *context*, *source*, and *excerpt* allows analysis of how evidence is introduced, cited, and used strategically during argumentation. Structured *annotations* of evidential quality and argumentative role enable investigation of which properties of evidence, such as *credibility*, *relevance*, *warrant strength*, or *offensive and defensive value*, most strongly influence judgment. Collectively, these variables make it possible to trace how evidence contributes to persuasion and adjudication decisions.

Adjudication Traces Adjudication traces are tuples consisting of the *adjudication* and the *justification* because they collectively reveal how decisions were reached (Zhang et al., 2016; Greschner et al., 2025; Tiwari et al., 2025). Capturing both the final judgment and its justification enables modeling of the reasoning process behind decisions across different debate formats, from audience vote shifts to expert scoring and rubric-based evaluations. Recording these traces makes adjudication decisions auditable, reproducible, and learnable for adjudication modeling.

3 Data Collection Protocols

This section outlines practical mechanisms for collecting debate data according to our proposed approach. Debates may include humans and/or AI (i.e., LLM) participants. We describe how debate episodes should be recorded across different participant configurations (H–H, H–LLM, and LLM–LLM) and how judges' decisions and reasoning should be captured to support downstream analysis of the adjudication process.

3.1 Debater-Generated Dialogue Data: Argument and case data collection

Human-Human Debate Considering that this blueprint can be used to guide novel data collections or compiling existing data sets, H-H debate transcripts can be assembled from recorded debate rounds, ongoing debate competitions, and original participant studies. The prompt of the debate and any inconsistencies with the structure (e.g. interruptions) (Loakes, 2024) and data provenance should be noted along with the transcription (Chandrasekar et al., 2024). The use of crowdsourcing is a potential method for recruiting participants to perform this annotation (Bhuiyan et al., 2020). If a new H-H debate dataset is to be collected, especially oral debate data, precise transcription tools should be use to ensure representativeness of vernaculars and accents of participants (Zolnoori et al., 2024; Point & Baruch, 2023), as such information may be useful for further analysis or evidence annotation, especially as related to adjudication. Of course, data collection should be subject to a participant's informed consent (Shaw et al., 2025).

Human-LLM Debate The collection of H-LLM dialogue data is similar to the H-H, but with additional logistics that need a formal structure (Mirkin et al., 2018). For example, the modality with which the LLM converses with the human should be documented. In the IBM Debater dataset for example, they use text-to-audio technology in an attempt to better mimic H-H debate (Slonim et al., 2021). Another approach would be a completely text-based debate, in which case design decisions, such as setting an LLM word limit to better mimic humans (Slonim et al., 2021), should be documented. Critically, human participants must know they are debating an LLM and give "genuinely informed consent" to ensure transparency and safeguard the human debaters (McKee, 2024), (nat, 2021).

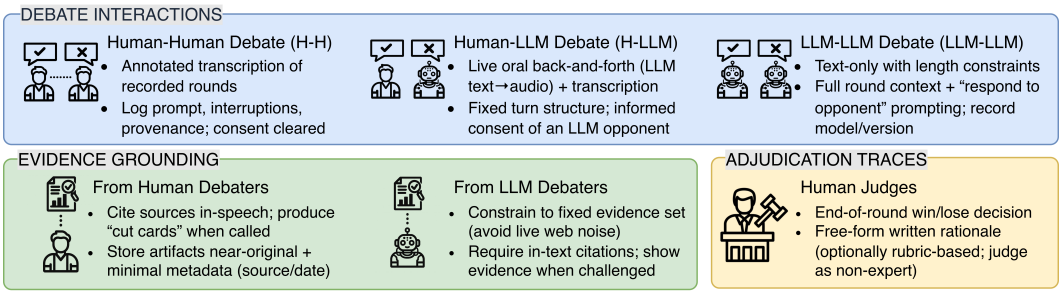

Figure 2: Guidelines for collecting debate adjudication data, spanning debate interactions, evidence grounding from human and LLM debaters, and human judge adjudication traces.

LLM-LLM Debate LLM-LLM debate can be generally limited to text-only dialogues. Given the rapidly evolving landscape of LLMs and their capabilities, what and how models is valuable information for both analysis and reproducibility purposes. For example, current LLMs are prone to providing run-on or unstructured arguments (Reinhart et al., 2025). This issue can be mitigated through prompting (Chuang et al., 2024) (e.g. by prompting the LLM to directly respond to the opponent's argument); such prompts should be recorded within the dataset if possible. While there is a practical concern that LLM-LLM datasets may lead to an overpopulation of low-quality synthetic data (Chim et al., 2025; Seddik et al., 2024), careful collection of the processes used to produce the data can mitigate those risks.

3.2 Collecting evidence compiled by debaters

Evidence grounding data should be obtained by explicitly separating evidence provided by human and LLM debaters; collection mechanisms therefore differ.

Evidence from human debaters Following the rule from formal H-H competitive debates, participants must ground their arguments in evidence and cite them throughout their speeches as they are being referenced (Snider & Schnurer, 2006). When requested ("called"), debaters must be able to present "cut cards" with the cited artifacts (e.g. documents, links, or prepared excerpts) stored as close to their original form, along with minimal provenance metadata such as source and date. Such evidence relies on what humans choose to provide or cite throughout the case (Roush et al., 2024; Snider & Schnurer, 2006). When it comes to data collection, regardless of an in-round "call," debaters should provide their true evidence.

Evidence from LLM debaters With LLMs, the approach to evidence collection and disclosure is less formalized and thus can present space for valuable discussions in the next step. Newest LLM versions have live access to Internet search, meaning they can have immense advantage over human debaters. To avoid "web-induced noise," it's best to constrain for human and LLM debaters to fixed evidence set (Wang et al., 2025a). To best mimic human debaters, LLMs should not be prompted to self-report their evidence but rather cite them when appropriate and disclose when prompted by the opponent (Roush et al., 2024).

3.3 Collecting adjudication traces

As previously mentioned, we are primarily concerned with human judges; as such, the collection of adjudication traces is relatively straightforward. The judge should report a win/lose outcome at the end of the round along with a natural language reasoning behind their decision, ideally referring to a rubric or overall performance. In formal debate competitions, judges are instructed to adjudicate without relying on topic expertise and instead evaluate arguments solely on the presented evidence (Khan et al., 2024). Nevertheless, judges' expertise should still be recorded to aid post-hoc analysis.

4 Downstream Research Applications

By centering debate episodes with dialogue, evidence grounding, and adjudication traces, the dataset schema enables mechanistic analyses of variables of interest and their effects on human judgment. To that end, we outline several research directions we plan to pursue in the future.

Table 1: Example research directions for modelling adjudication biases using the proposed variables.

Cognitive Bias	Debate Dialogue	Evidence Grounding	Adjudication Traces
Anchoring	`stage`, `turn_id`	`speech_position`, `best_use_timing`	`reason`
Halo Effect	`speaker_id`, `personnel_id`	`author_credibility`, `citation_quality`, `technical_complexity`	`reason`
Confirmation Bias	`content`, `paragraph_id`	`debate_relevance`, `argument_type`, `counter_argument_resistance`	`reason`
Availability Heuristic	`content`	`evidence_type`, `jargon_density`, `scope_of_impact`	`reason`
Framing Effect	`content`, `paragraph_id`	`argument_type`, `offensive_value`, `defensive_value`, `warrant_strength`	`reason`

4.1 Behavioral Analyses of debate dialogue

Due to the granularity of its turn-structured dialogue, the dataset schema yields metadata about the relation between potential variables of interest. We offer the domain of biases, defined as systematic, unconscious factors that affect decision-making, as one possible lens through which these variables could be analyzed (Kahneman et al., 1982). Since we focus on modeling the human adjudicator, we are especially interested in identifying cognitive biases as hidden features of debate datasets. In some cases, these biases may be measured directly through observable signals, but in many settings they remain partially implicit, shaping judgments in ways that require inference in place of measurement. Drawing inspiration from work that treats human-centered attributes as latent structure in sequential decision-making models (e.g., Chen et al., 2018), we intend to use attributes of interest, such as persuasion, honesty, and verbosity, as latent variables for fitting a model in the debate setting.

We offer an example set of human biases that could be studied in subsequent work (Table 1), drawing from the cognitive biases specified in Malberg et al. (2025) and presenting them as proposed latent features that could be derived from a structured debate dataset. Due to their relevancy to the structural and cognitive elements of the debate setting, we propose to study five biases: the halo effect, confirmation bias, availability heuristic, and framing effect. The cognitive bias of anchoring, for example, refers to the tendency to emphasize the first piece of information received in decision-making processes, and can be derived using sequential attributes such as `turn_id` and `speech_position` (Rezaei, 2021).

Though we foreground the biases of the human adjudicator, it is important to note that the growing class of LLM-specific biases can also affect the outcome of a debate. LLM biases include but are not limited to egocentric biases (Panickssery et al., 2024), position bias (Wang et al., 2025b), and sycophancy (Malmqvist, 2025), some of which have parallels in humans but all of which can skew their behavior in a debate setting.

4.2 The training and evaluation of robust debate adjudication protocols

In addition to analysis, well-structured datasets advance debate as a training and evaluation paradigm (Slonim et al. (2021). By identifying the attributes that make debates persuasive to human evaluators, protocol developers can account for cognitive biases and train debaters to be robust to realistic judge models. Moreover, reconstructing and auditing single reference frames and entire debate progression allows for automated checks of evidence faithfulness and citation quality. Future dataset schema can further adapt adjudication criteria that are poorly specified and integrate attributes that are missing altogether.

5 Conclusion

Existing debate datasets rarely capture in one place the dialogue dynamics, evidential support, and reasoning traces that shape how judges reach decisions, but these components are critical for understanding the process of adjudication. In this preliminary work, we presented a blueprint for collecting structured debate adjudication datasets that couple turn-level interaction, evidence grounding, and adjudication outcomes within a unified schema, alongside practical considerations for debate data collection. By treating adjudication as a central empirical object of study, this framework enables downstream work on modeling judge behavior, auditing persuasion and evidence use, and designing debate-based evaluation and training methods that account for systematic biases rather than assuming idealized supervision by oracles.

REFERENCES

Am i arguing with a machine? AI debaters highlight need for transparency. *Nature*, 592:166, April 2021. doi: 10.1038/d41586-021-00867-6. URL https://doi.org/10.1038/d41586-021-00867-6. Editorial.

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4445–4452, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1704/.

Md Momen Bhuiyan, Amy X. Zhang, Connie Moon Sehat, and Tanushree Mitra. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2), October 2020. doi: 10.1145/3415164. URL https://doi.org/10.1145/3415164.

Abinaya Chandrasekar, Sigrún Eyrúnardóttir Clark, Sam Martin, Samantha Vanderslott, Elaine C. Flores, David Aceituno, Phoebe Barnett, Cecilia Vindrola-Padros, and Norha Vera San Juan. Making the most of big qualitative datasets: a living systematic review of analysis methods. *Frontiers in Big Data*, Volume 7 - 2024, 2024. ISSN 2624-909X. doi: 10.3389/fdata.2024.1455399. URL https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1455399.

Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, pp. 307–315, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450349536. doi: 10.1145/3171221.3171264. URL https://doi.org/10.1145/3171221.3171264.

Jenny Chim, Julia Ive, and Maria Liakata. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, 51(1):191–233, March 2025. doi: 10.1162/coli_a_00540. URL https://aclanthology.org/2025.cl-1.6/.

Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Beyond demographics: Aligning role-playing llm-based agents using human belief networks, 2024. URL https://arxiv.org/abs/2406.17232.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

Lynn Greschner, Sabine Weber, and Roman Klinger. Trust me, i can convince you: The contextualized argument appraisal framework, 2025. URL https://arxiv.org/abs/2509.17844.

Chris Guthrie, Jeffrey J. Rachlinski, and Andrew J. Wistrich. Judging by heuristic: Cognitive illusions in judicial decision making. *Judicature*, 86(1):44, July-August 2002. URL https://scholarship.law.cornell.edu/facpub/862.

Chris Guthrie, Jeffrey J. Rachlinski, and Andrew J. Wistrich. Blinking on the bench: How judges decide cases. *Cornell Law Review*, 93:1, 2007. URL https://scholarship.law.cornell.edu/facpub/917.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate, 2018. URL https://arxiv.org/abs/1805.00899.

Daniel Kahneman, Paul Slovic, and Amos Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press, Cambridge, 1982. doi: 10.1017/CBO9780511809477.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers, 2024. URL https://arxiv.org/abs/2402.06782.

Debbie Loakes. Automatic speech recognition and the transcription of indistinct forensic audio: how do the new generation of systems fare? *Frontiers in Communication*, Volume 9 - 2024, 2024. ISSN 2297-900X. doi: 10.3389/fcomm.2024.1281407. URL https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2024.1281407.

Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in LLMs. In Mika Hämäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar (eds.), *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pp. 578–613, Albuquerque, USA, May 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.nlp4dh-1.50. URL https://aclanthology.org/2025.nlp4dh-1.50/.

Lars Malmqvist. Sycophancy in large language models: Causes and mitigations. In Kohei Arai (ed.), *Intelligent Computing*, volume 1426 of *Lecture Notes in Networks and Systems*, pp. 61–74. Springer, Cham, 2025. doi: 10.1007/978-3-031-92611-2_5. CompCom 2025.

Kevin R. McKee. Human participants in ai research: Ethics and transparency in practice. *IEEE Transactions on Technology and Society*, 5(3):279–288, September 2024. ISSN 2637-6415. doi: 10.1109/tts.2024.3446183. URL http://dx.doi.org/10.1109/TTS.2024.3446183.

Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian, and Noam Slonim. A recorded debating dataset, 2018. URL https://arxiv.org/abs/1709.06438.

Ted Moskovitz, Aaditya K. Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D. Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained rlhf, 2023. URL https://arxiv.org/abs/2310.04373.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations, 2024. URL https://arxiv.org/abs/2404.13076.

Sébastien Point and Yehuda Baruch. (re)thinking transcription strategies: Current challenges and future research directions. *Scandinavian Journal of Management*, 39(2):101272, 2023. ISSN 0956-5221. doi: https://doi.org/10.1016/j.scaman.2023.101272. URL https://www.sciencedirect.com/science/article/pii/S0956522123000131.

A. Reinhart, B. Markey, M. Laudenbach, K. Pantusen, R. Yurko, G. Weinberg, and D. W. Brown. Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences of the United States of America*, 122(8):e2422455122, 2025. doi: 10.1073/pnas.2422455122.

Jafar Rezaei. Anchoring bias in eliciting attribute weights and values in multi-attribute decision-making. *Journal of Decision Systems*, 30(1):72–96, 2021. doi: 10.1080/12460125.2020.1840705. URL https://doi.org/10.1080/12460125.2020.1840705.

Allen G Roush, Yusuf Shabazz, Arvind Balaji, Peter Zhang, Stefano Mezza, Markus Zhang, Sanjay Basu, Sriram Vishwanath, and Ravid Shwartz-Ziv. Opendebateevidence: A massive-scale argument mining and summarization dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=43s8hgGTOX.

Ramon Ruiz-Dolz and Javier Iranzo-Sánchez. Vivesdebate-speech: A corpus of spoken argumentation to leverage audio features for argument mining, 2024. URL https://arxiv.org/abs/2302.12584.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. How bad is training on synthetic data? a statistical analysis of language model collapse, 2024. URL https://arxiv.org/abs/2404.05090.

Heather Shaw, Olivia Brown, Joanne Hinds, Sophie J. Nightingale, John Towse, and David A. Ellis. The decide framework: Describing ethical choices in digital-behavioral-data explorations. *Advances in Methods and Practices in Psychological Science*, 8(3):25152459251361013, 2025. doi: 10.1177/25152459251361013. URL https://doi.org/10.1177/25152459251361013.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. An autonomous debating system. *Nature*, 591(7850):379–384, 2021. URL https://doi.org/10.1038/s41586-021-03215-w.

Alfred Snider and Maxwell Schnurer. *Many Sides: Debate Across the Curriculum*. International Debate Education Association, revised edition, July 2006. URL https://www.uvm.edu/~debate/dcpdf/Many_Sides_2nd_ed.pdf. 2nd edition.

Utkarsh Tiwari, Aryan Seth, Adi Mukherjee, Kaavya Mer, Kavish, and Dhruv Kumar. Debatebench: A challenging long context reasoning benchmark for large language models, 2025. URL https://arxiv.org/abs/2502.06279.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence, 2025a. URL https://arxiv.org/abs/2504.13079.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. 2025b. URL https://arxiv.org/abs/2407.01100.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational flow in oxford-style debates, 2016. URL https://arxiv.org/abs/1604.03114.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

Maryam Zolnoori, Sasha Vergez, Zidu Xu, Elyas Esmaeili, Ali Zolnour, Krystal Anne Briggs, Jihye Kim Scroggins, Seyed Farid Hosseini Ebrahimabad, James M Noble, Maxim Topaz, Suzanne Bakken, Kathryn H Bowles, Ian Spens, Nicole Onorato, Sridevi Sridharan, and Margaret V McDonald. Decoding disparities: evaluating automatic speech recognition system performance in transcribing black and white patient verbal communication with nurses in home healthcare. *JAMIA Open*, 7(4):ooae130, 12 2024. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooae130. URL https://pubmed.ncbi.nlm.nih.gov/39659993/.